EXPRESSION PROFILE ANALYSIS
AND
VALIDATION OF GENES ASSOCIATED WITH BREAST CANCER

Dissertation submitted in partial fulfillment of the requirement for the degree of Master of Science in Life science

BY:

BIBHUDUTTA MISHRA

ROLL NO: 411LS2051

SUPERVISED BY:

DR. BIBEKANAND MALLICK

DEPARTMENT OF LIFE SCIENCE

NATIONAL INSTITUTE OF TECHNOLOGY

ROURKELA-769008

NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA
राष्ट्रीय प्रौद्योगिकी संस्थान, राउरकेला

Dr. Bibekanand Mallick, **M.Tech., Ph.D.**
Assistant Professor

RNA Biology & Functional Genomics Lab.
Department of Life Science
National Institute of Technology
(*Ministry of H.R.D, Govt. Of India*)
Rourkela - 769 008, Odisha, India

Telephone: +91-661-246 2685 (O)
E-mails: vivek.iitian@gmail.com, mallickb@nitrkl.ac.in
Homepage: http://vvekslab.in

Date: 10. 05. 2013

CERTIFICATE

This is to certify that the thesis entitled ***"EXPRESSION PROFILE ANALYSIS AND VALIDATION OF GENES ASSOCIATED WITH BREAST CANCER"*** *submitted to National Institute of Technology; Rourkela for the partial fulfillment of the Master degree in Life science is a faithful record of bonafide and original research work carried out by* ***BIBHUDUTTA MISHRA*** *under my supervision and guidance.*

(Dr. Bibekanand Mallick)

DEDICATED TO

MY FAMILY AND BELOVED

ONES.....

ACKNOWLEDGEMENT

I express my deep sense of gratitude and reverence to my supervisor, Dr. Bibekanand Mallick, Department of Life Science, NIT Rourkela for his excellent guidance, constant and untiring supervision, help and encouragement throughout the investigation and preparation of the thesis.

I am extremely grateful and indebted to Dr. S.K. Patra, HOD, Department of Life Sciences, NIT, Rourkela and other faculty members, Dr. S.K. Bhutia, Dr. Bismita Nayak, Dr. S. Das, Dr. Rasu Jayabalan and Dr. Suman Jha for their inspiring suggestions and valuable advice during the course of studies and projects.

I am highly obliged to Devyani Samantarrai, Debashree Das, Dipto Sengupta, Pradipta ranjan Routa, Durgesh Nandini Das and Hirak Ranjan Dash, Ph.D. Scholars of the Department of Life Science, NIT-Rourkela for their constant help and encouragement during the period of my project. I am solely impressed by their great personalities.

My heartfelt thanks to my friends Sandeep Dey, Krishna Mohanty, Bini Chetri, MD Khurshid Ul Hassan, Mitali Rana, Subhrata Jena, Dibyojyoti Baruah, Tapas Tripathy and all other classmates for their moral support, help and encouragement throughout the course of this work. I take the pleasure to acknowledge the constant help and support of my friends has always been cherished.

My sincere obligations are to Mr. B. Das and Murali Mausa, Staffs of Department of Life Sciences, NIT, Rourkela for their help during this period.

Lastly, I acknowledge with highest sense of regards to my family for their supreme sacrifice, blessings, unwavering support, love and affection without which the parent investigation would not have been successful in any sphere of my life.

At the end, I bow down my head to the almighty whose omnipresence has always guided me and made me energized to carry out such a project.

Date: **BIBHUDUTTA MISHRA**

Place:

Contents

List of tables

List of figures

	parameter analysis
17	Interaction of EGR1 gene with other gene by taking Experiment-Text mining parameter analysis
18	Interaction of EGR1 gene with other gene by taking Database-Text mining parameter analysis
19	Interaction of LAMB1 gene with other gene by taking Experiment-Database-Text mining parameter analysis
20	Interaction of LAMB1 gene with other gene by taking Text mining parameter analysis
21	Interaction of LAMB1 gene with other gene by taking Experiment-Text mining parameter analysis
22	Interaction of LAMB1 gene with other gene by taking Database-Text mining parameter analysis
23	Egr1showing Association with its neighboring gene
24	Lamb1showing Association with its neighboring gene
25	Results of melting curve analysis of both EGR1 and LAMB1 gene using β actin as standard
26	Results of melting curve analysis of EGR1 gene using β actin as standard
27	Results of melting curve analysis of LAMB1 gene using β actin as standard
28	Relative expression of EGR1 AND LAMB1 with respect to control

Abstract

ABSTRACT

Breast cancer is one of the most common causes of cancer related death in women around the globe. The process of tumor invasion and subsequent metastasis represents the most lethal aspect of breast cancer like any other cancer and is responsible for the majority of deaths among cancer patients. Our objective was to identify specific genes involved in such phenomenon or any other mechanism/networks in breast cancer. We analyzed microarray data of breast cancer tissues samples and compared with array of normal breast epithelium to obtain a list of 255 genes that are differentially expressed in breast cancer patients. Out of 255, 153 genes were reported to be down-regulated and 102 genes were up-regulated in breast cancer. From this differentially expressed genes list, we choose two genes, EGR1 & LAMB1 for experimental validation in cell lines, because their de-regulation might be contributing factor of breast cancer by affecting pathways/processes in the body. The expressions of these two genes were validated by qRT-PCR in MDA-MB-231 breast cancer cell lines, followed by elucidating their association in various biological networks.

Key words: Microarray, clustering, breast cancer, qRT-PCR, cancer cell lines

Introduction

1. INTRODUCTION

The term cancer was coined by Hippocrates in the fifth century BC, which means "crab" in Latin. It describes a condition in which cells divide and spread unrestrained throughout the body, eventually choking off life. The disease has existed for at least several thousand years and its prevalence has been steadily increasing. It is seen that cancer strikes older people more frequently than younger people and more cancer cases are being seen in older age. The increase in average lifespan-due mainly to the availability of vaccines and antibiotics that have lowered death rates from infectious diseases-resulting in more and more people living long enough to develop cancer.

Breast cancer is one of the most common causes of cancer related death in women over the world. Globally, breast cancer accounts for an estimated 1.4 million cases each year, with more than half of the 400,000 breast cancer deaths occurring in low and middle income countries. Although therapeutic approaches, such as surgery, chemotherapy and radiation therapy, have reduced cancer specific mortality, there still are many therapeutic failures which result in cancer recurrence, metastasis and death.

Gender and age are important risk factors for breast cancer; rates are about 100 times higher in women than in man, and about 80% of all cases are diagnosed in women over 50 years of age. Roughly 10% of all breast cancers are linked to a family history of a disease, associated mainly with inherited mutation in the BRCA1 or BRCA2 genes. Those who have had extensive exposure to ionizing radiation in the chest area incur an increased risk for breast cancer, and reproductive and hormonal history play an important role as well. It is also seen that menstruation at an early age or go through menopause at a late age are at slightly increased risk, as are women who have had no children or who had their 1st child after age 30. From a wide range of study it was seen those who used oral contraceptives; exhibit a slight elevation in breast cancer risk. Use of alcohol, obesity and lack of physical exercise are additional risk factors. Some reports have suggested that environmental pollution involving pesticides that mimic the action of estrogens can promote the development of breast cancer, although current evidence does not show a consistent link.

The 1st symptom of breast cancer is usually a lump detected during breast self-examination. Unexplained breast swelling, thickening, skin irritation or dimpling, tenderness, nipple pain and discharge other than breast milk are other possible signs. Mammography is capable of detecting small breast cancers before they can be felt by a woman or by her

physician. It is therefore currently being recommended that women over age 40 begin regular mammography to screen for breast cancer. When physical examination or mammography indicates the presence of an abnormal mass, a biopsy is done to determine whether cancer is present.

Breast cancer has been attributed to diverse causes but the real mechanism underlying cancer development remains obscure. Elucidation of such mechanism requires thorough knowledge of the genes involved and their regulation strategies. With the invention of DNA microarray technology, it is now possible to look globally at gene expression across the genome. By studying gene expression profile we can know the interactions between genes and how this ultimately impacts on the pathology of diseases. (Fey M et. al. 2002). Such DNA microarray analysis has been used to investigate the underlying biology of many cancer types, including breast cancer **(**Duffy et .al**,** 2005).This type of analysis can be applied to every aspect of the disease including initiation, progression, invasion and drug resistance with the aim of elucidating the underlying molecular mechanisms involved (Bertouchi et.al.,2001). There have been many studies using breast cancer cell lines (Zajchowski and Perou et.al , 2000) which have used DNA microarrays in an attempt to develop molecular portraits, classifications and prognostic signatures of breast cancer (Brennan et.al., 2005).

The main aim of our work is to find a set of key genes which have significant roles in breast cancer like tumor suppression and inhibiting metastasis and also to find out how they are regulating other genes or pathways. If we somehow regulate their expression, there might be a cure or reduction in tumor invasiveness. To do that, microarray analysis of tumor versus control sample was done using Genespring, followed by gene ontology study using String and Genomatix software. These analysis lead to the selection of two genes for further study based upon their regulation and association with cancer. Expressions of these two genes were confirmed in MDA-MB-231 cell line which is a breast cancer cell line, taking β-actin (a housekeeping gene) as reference gene through qRT-PCR. This study can lead to possible discovery of genes which play a key role in tumor invasiveness.

Review of literature

2. REVIEW OF LITERATURE

Cancer is characterized by the abnormal growth and proliferation of cells in an uncontrolled manner .Such cells can arise in a variety of tissues and organs and each of these sites contains different cell types.

Benign and malignant tumours

No matter where a cancer arises and regardless of the cell type involved, the diseased can be defined by a combination of 2 properties i.e.

- The ability of the cells to proliferate in an uncontrolled fashion.
- Their ability to spread throughout the body.

Tumours types

On the basis of differences in the growth patterns of tumours are subdivided in to 2 fundamentally different categories. One group consists of **benign tumors**, which grow in a confined local area. Another is **malignant tumors** that can invade surrounding tissues, enter the bloodstream, and spread to distant parts of the body by the process called **metastasis**. The cells of a malignant tumour often spread to other parts of the body. Malignant tumours therefore tend to be more hazardous than benign ones. Benign tumours on the other hand arise in surgically inaccessible locations, such as the brain, making them hazardous and potentially life threatening.

Differences in growth rate and state of differentiation are also common between benign and malignant tumours. Benign tumours generally grow rather slowly and are composed of well differentiated cells, meaning that the cells bear a close structural and functional resemblance to the normal cells of the tissue in which the tumour has arisen. Malignant tumours, on the other hand ,often grow more rapidly and their state of differentiation is variable, ranging from relatively well differentiated tumours to tumours whose cells are so poorly differentiated that they bear almost no resemblance to the original cells from which they were derived. Therefore the malignant cell pool constitutes a heterogeneous population of cells with varying degrees of differentiation, from stem cell like to fully differentiated ones.

Nomenclature of tumors according to site of origin

Because tumours can arise from a variety of cell types located in different tissues and organs, some basic conventions have been established to facilitate the naming of cancers. Depending on their site of origin, cancers are grouped into 3 main categories i.e.

(1) **Carcinomas** are cancers that arise from the epithelial cells that form covering layers over external and internal body surfaces. Carcinomas are by far the most common type of malignant tumour, accounting for roughly 90% of all human cancers.

(2) **Sarcomas** are cancers that originate in supporting tissues such as bone, cartilage, blood vessels, fat, fibrous tissue and muscle. They are the rarest group of human cancers, accounting for about 1% of the total.

(3) The remaining cancers are **lymphomas** and **leukaemia,** which arise from cell of lymphatic and blood origin. The term lymphoma refers to tumours of lymphocytes that grow mainly as solid masses of tissue, whereas leukaemia are cancers in which malignant blood cells proliferate mainly in the bloodstream(Kleinsmith et.al.,2009) .

Table 1: Nomenclature of tumors according to their origin (Kleinsmith et.al. 2009)

Prefix	**Cell type**	**Benign tumor**	**Malignant tumor**
Epithelium			
Adeno	Gland	Adenoma	Adenocarcinoma
Basal cell	Basal cell	Basal cell adenoma	Basal cell carcinoma
Squamous cell	Squamous cell	Karatoacanthoma	Squamous cell carcinoma
Melano	Pigmented cell	Mole	Melanoma
Supporting Tissue			
Chondro	Cartilage	Chondroma	Chondrosarcoma
Fibro	Fibroblast	Fibroma	Fibrosarcoma
Hamangio	Blood vessels	Hemangioma	Hemangiosarcoma
Leiomyo	Smooth muscle	Leiomyoma	Leiomyosarcoma
Lipo	Fat	Lipoma	Liposarcoma
Meningio	Meninges	Meningioma	Meningiosarcoma
Myo	Muscle	Myoma	Myosarcoma
Osteo	Bone	Osteoma	Osteosarcoma

Rhabdomyo	Striated muscle	Rhabdomyoma	Rhabdomyosarcoma
Blood and lymph			
Lympho	Lymphocyte		Lymphoma
Erythro	Erythrocyte		Erythrocytic leukemia
Myelo	Bone marrow		Myeloma

Breast cancer has been attributed to diverse causes but the real mechanism underlying cancer development remains obscure. The cause lies underneath, may be due to activation or inhibition of different regulatory pathways or it may be due to up or down regulation of some genes. To solve the mystery an effective technology is being used known as DNA microarray.

DNA microarray analysis is a powerful tool in the armory of the molecular biologist, as it allows examination of thousands of genes in parallel (Schena et.al., 1998). With the invention of DNA microarray technology, it is now possible to analyze gene expression across the genome simultaneously, under identical conditions and to find patterns of expression. Patterns of gene expression should lead to a better understanding of the complex interactions between genes and how this ultimately impacts on the pathology of disease (Fey M et.al, 2002).

DNA microarray analysis has been used to investigate the underlying biology of many cancer types, including breast cancer (Brennan, et.al.2005). This type of analysis can be applied to every aspect of the disease including initiation, progression, invasion and drug resistance with the aim of elucidating the underlying molecular mechanisms involved (Bertouchi et.al. 2001). There are many published studies using breast cancer cell lines (Zajchowski et.al., 2001)and breast tumor tissue (Perou et.al.,2000) which have used DNA microarrays in an attempt to develop molecular portraits, classifications and prognostic signatures of breast cancer (Brennan, et.al.2005). An example of the potential for DNA microarrays in the clinical management of patients with cancer is the 70-gene prognostic breast cancer signature (Van't Veer et.al. 2002), which is now undergoing evaluation in clinical trials. Using of this technology for the characterization of a series of invasive subclones with different levels of invasion in primary human breast cancer cell line, Hs578T, is being established.

Different tools and softwares are used to analyze the microarray data and set up biological networks. It includes:

- **Genespring**: It is a commercially available software used for microarray analysis.
- **String:** It is a database to set up interaction between different genes.
- **Genomatix:** It is a commercially available software for establishing biological network and pathway interaction study.
- **Bioconductor:** It is an open source software for analyzing microarray data.
- **TM4**: It is a free software for microarray analysis and it has mainly 4 applications i.e. Spotfinder, Microarray Data Manager (MADAM),Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV).
- **Spotfire:** It is commercially available software for microarray data analysis.
- **R**: It is free software for statistical computing and graphics.

Objectives

OBJECTIVES

Objective 1

- To check the expression profile analysis of genes associated with breast cancer using microarray data.

Objective 2

- Selection of a set of key genes associated in breast cancer.

Objective 3

- Experimental validation of selected genes.

Materials and Methods

3. MATERIALS & METHODS

Gene Expression data

First we have to check mRNA expression between normal vs. diseased. To see that following approach is taken.

- To see mRNA expression between normal vs. diseased DNA microarray data is used from the GEO database.
- The DNA microarray data are stored in GEO database. So from there only we can retrieve this type of data.

3.1 Retrieval of gene expression data

- The gene expression data was retrieved by using of GEO database.

GEO database

GEO (http://www.ncbi.nlm.nih.gov/geo/) is a data repository and retrieval system for any high throughput gene expression. It contains data from different microarray experiments as well as non-array based technologies such as SAGE and mass spectrometry peptide profiling. GEO segregates data in to 3 principle components, platform, sample and series. A platform is essentially a list of probes that define what set of molecules may be detected. A sample is a set of molecules that are being probed and references a single platform used to generate its molecular abundance data. A series organizes samples in to meaningful datasets which make up an experiment (Mallick et.al. 2012).

Basic retrieval of GEO data

There are 2 ways by which the GEO data may be retrieved

- GEO data may be retrieved by querying GDS (GEO datasets), gene profiles and GEO accession numbers.
- GEO data can be accessed directly on the web by browsing through datasets or GEO accessions options available with respect to individual platform, sample and series. Related records are intra-linked on the GEO sites such that one may conveniently navigate to the associated platform, , series, sample and GEO dataset records.

Analysis through GEO data base

- Go to main page of GEO data base.
- Type mRNA expression in breast cancer AND homo sapiens in data setsoption.
- Then click on GO.

Figure 1: Retrieval of microarray data from the GEO database for the experiment

- Then results appear which contain the list of microarray experiments.
- From that each GSE is checked for expression profile analysis having platform [HG-U133A] Affymetrix Human Genome U133A Array and the expression is without induce of any drug.
- From that **GSE9574 is** taken for analysis.
- From that 3 controls (normal) and 3 tests (disease) samples were taken.
- For control GSM242005
  GSM242006
  GSM242007 is taken for analysis.
- For test GSM242020
  GSM242021
  GSM242022 is taken for analysis.
- It was taken in triplicate because to minimize the error.
- From that a raw file is down loaded, it was unzip and for further analysis it is uploaded in GeneSpring software.

- To see which genes are associated with the expression of mRNA and their regulation in breast cancer software is used known as GeneSpring.

3.2 Analysis of gene expression data

- Expression data is analyzed by using Genespring software.

GeneSpring software

GeneSpring is software which provides statistical tools for visualization and data analysis. It helps to investigate and understanding of Transcriptomic, Metabolomics, Proteomics and NGS data within a biological context. It also helps in expression analysis. This software is a key component of systems biology research involves producing heterogeneous data that measure various biological entities and events such as variation in DNA structure, expression of microRNA and mRNA exon splicing, proteins and metabolites, which helps to understand underlying mechanism of diseases.

It also allows researchers to analyse, compare different signalling pathway.

Other applications of GeneSpring include:

- Gene expression analysis of microarray platforms like Affymetrix, Agilent and Illumina.
- Analysis of Real-time PCR data.
- GEO datasets.

Figure 2: Clustering analysis of normal and diseased samples

Data analysis using GeneSpring

For mRNA expression analysis:

Procedure

- The raw data files from GEO database were downloaded as a zip file.
- Then the files were unzipped, extracted and renamed according to the convenience.
- First we have to select a new project under which many platforms will come from that we have to select our desired platform type.
- Then we have to give the name of the project.
- Then we have to set a new experiment by setting experiment name and experiment type.
- Then guided work flow is chosen as work flow type.
- Then data is loaded and the desired technology is selected i.e. Affymetrix Gene Chip-HG-U113A.
- Then files were chosen as 3 replicates i.e. Control and Test files.
- Then by normalizing the noise signal the intensity map can be seen.
- Then **Experimental setup** was done by following methods.

❖ First **experimental grouping** is done by add parameters to **Average.**

❖ Then **Average over replicates in condition** is chosen in case of **Create Interpretation.**

- Then **quality control** is done by following methods.

❖ First we have to set **Filter Probe sets by errors.**

❖ Then the **coefficient of variation** is chosen **< 50%.**

❖ Then we have to choose **Output views of filter.**

❖ Then the **entity list** is saved.

- Then **Analysis** was done by following method

❖ First **Statistical Analysis** is chosen.

❖ After that **Entity list** is Filtered on **error <50.**

❖ Then the **Interpretation** is chosen **Average**.

❖ **Condition 1** and **Condition 2** for **Select test**.

❖ **T-test is unpaired chosen** for 2 sets of data. If more than 2 sets be present then **Anova** should be chosen.

- ❖ Then the **p- value computation** is chosen **Asymptomatic.**
- ❖ Then the **Number of permutation** should be **100**.
- ❖ Then **Multiple testing correction** is chosen **Benjamin Hochberg FDR.**
- ❖ Then in **Results** cut-off should be chosen **≤ 0.05** which is present by default.
- ❖ Then the **entity list** is saved.
- Then the **Fold Change** is chosen and done by following these **Input parameters** and **pairing options.**
- ❖ The **Fold change results** cut-off should be **>=2.0.**
- Then **Clustering** was done by following methods:
- ❖ First **Input parameters** are taken.
- Then in **Entity list**, Fold change **>=2.0** is chosen**.**
- Then the **Interpretation** is chosen **all samples**.
- The **Cluster algorithm** is chosen **Hierarchical.**
- The **Cluster** is done on **Conditions**
- Then the **Distance metric** is chosen **Pearson uncentered**.
- The **Linkage rule** is chosen Average**.**
- Then Result **file** is saved.
- After that **Result interpretation** was done by following method
- Then GO analysis by choosing output views.
- **Export** of **Fold change data** was done by adapting the method:
- ❖ First we have to **Export** the **entity list**.
- ❖ Then we have to normalize the **signal values.**
- ❖ Then - **Gene symbol** and **Entrez gene** is selected in **selected items**.
- ❖ Then **Interpretation** of **All samples** is done.
- Then it is Saved as- Tab separated file (.txt)

- Then we have to see how the control varies from test sample.
- To see how the control varies from test sample clustering is done by using **Cluster3** and **java treeview software** is used.

3.3.Cluster3 and java treeview software

Cluster3 and **java treeview** are the softwares which help in analyzing data from the **GeneSpring.**

Loading, filtering, and adjusting data

We can be load data into Cluster by choosing Load data file clicking the File menu. Then a number of options appear which helps in adjusting and filtering the loaded data. These can be done by Filter Data and Adjust Data tabs.

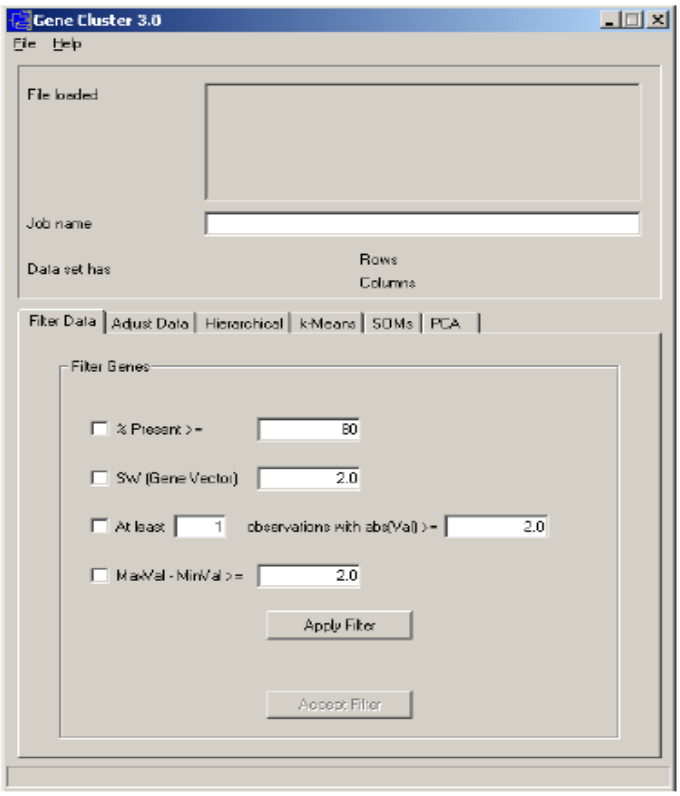


Figure 03: Analysis of data using cluster3 software

Loading Data

The first step for Cluster is to import data. The data can be imported from the Microsoft Excel file. Cluster only reads tab-delimited text files so we have to first convert it in to .txt format to upload. For e.g. a demo data file is uploaded for checking the expression of the normal and diseased samples.

Filtering Data

With the help of Filter Data tab one can remove genes that do not have desired properties which we need for our dataset.

Clustering techniques

The Cluster program provides several clustering algorithms. There are 4 clustering methods are available i.e. Hierarchical ,K-means, SOM(Self-Organizing Maps) and PCA(Principal Component Analysis).We mainly use Hierarchical clustering methods for our analysis. It organizes genes in a tree structure, based on their similarity.

Java treeview

Java treeview is software which allows analyzing the results of Cluster. It reads in .cdt, .gtr, .atr, .kgg, or .kag file formats which are being produced by Cluster.

- After that we have to see how these 2 genes regulate expression by associating with their neighboring genes.
- To see that another database is used known as string. (Search Tool for the Retrieval of Interacting Genes database).

3.4. String database

Proteins can form a variety of functional connections witheach other, including stable complexes, metabolic pathways and a bewildering array of direct and indirect regulatory interaction. Otherwise these can be known as networks and the size and complex organization of these networks present a unique opportunity to view a given genome as something more than just a static collection of distinct genetic functions. Indeed, the ‘network view’ on a genome is increasingly being taken in many areas of applied biology: protein networks are used to increase the statistical power in human genetics (Pattin, et.al., 2009; Pujol et.al., 2010) to close gaps in metabolic enzyme knowledge and to predict phenotypes and gene functions (Lage. et al, 2007; Wang et.al., 2010) to name just a few examples. Protein– protein association information is highly useful for the users and it should be easily available in web and should be easily accessible for the users. The Search Tool for the Retrieval of Interacting Genes (STRING) database resource aims to provide this service, by acting as a ‘one-stop shop’ for all information on functional links between proteins. STRING is the only site which covers the interaction map of hundreds more than 1100 organisms ranging from single cell organism to humans.

Analysis using string database

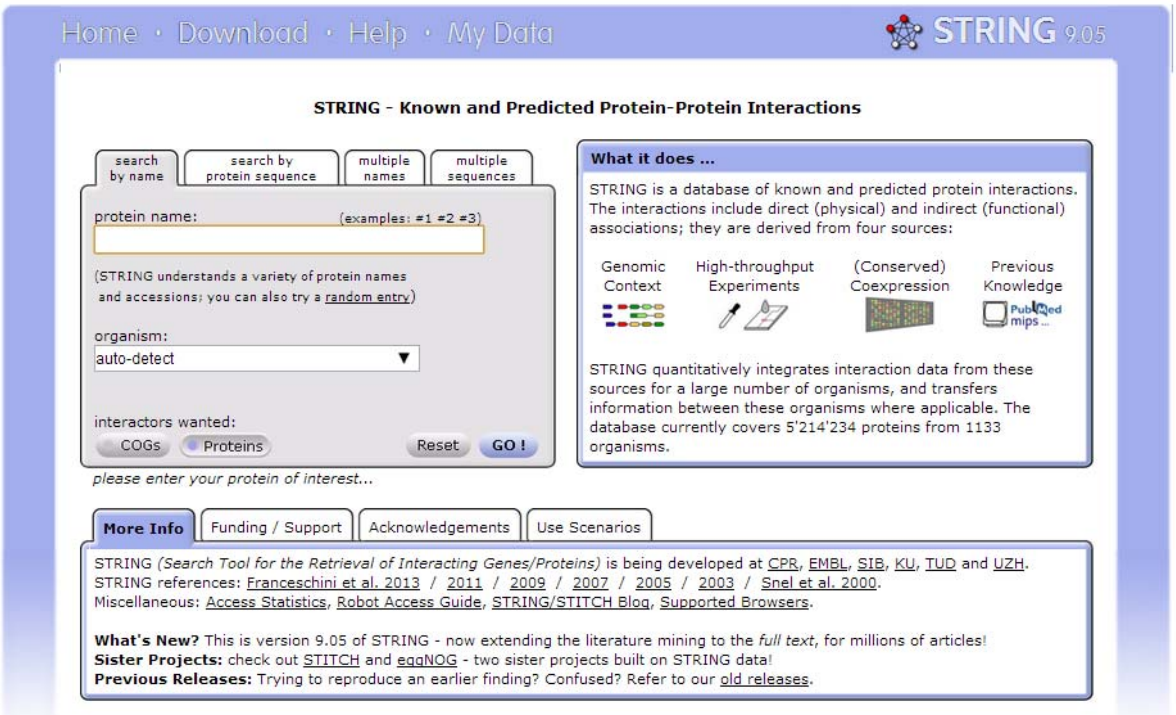


Figure 04: Analysis of genes using string database

- Here multiple names are selected.
- *homo sapiens* should be selected in organism name.
- After that gene list should be added in **list of name** box.
- Then **GO** option is chosen.
- Then it will show the gene list which are uploaded and their function**.**
- Then **continue** option is chosen for further analysis.
- By changing different parameter analysis is done.
- Then update parameters for the interaction analysis.

- Again it is analyzed in software named as **Genomatrix.**
- It is also interaction based software.

- Then the expression data obtained from the gene spring was further validated by wet lab analysis by selecting the 2 genes i.e.**EGR1** and **LAMB1** with a normal gene named as **β actin**.

3.5 Experimental Validation

3.5.1 Cell culture

- MDA- MB-231 breast cancer cell lines are cultured for isolation of RNA.

MDA MB 231 cell line

The MDA-MB-436 breast cancer cell line was first derived from pleural fluid obtained from a 43-year-old breast cancer patient in 1976. It is pleomorphic and reacts intensely with anti tubulin antibody. The MDA-MB-231 cell line is also able to grow on agarose, an indicator of transformation and tumorigenicity, and displays a relatively high colony forming efficiency. *In vivo*, the MDA-MB-231 cells form mammary fat pad tumors in nude mice.

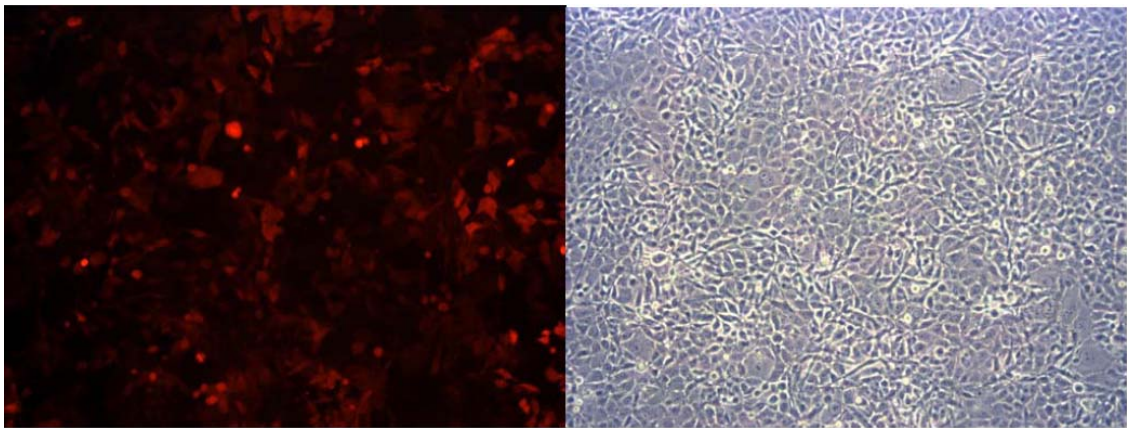


Figure 05. MDA-MB-231 Cell Line. Left: RFP Fluorescence; Right: Phase Contrast.

- Before culturing these cell lines are maintained in 20° c in carbon dioxide incubator.
- Media used for these cell cultures is **DMEM.**
- Freeze Medium is 70% DMEM, 20% FBS, 10% DMSO.

Methods of cell culture

Human breast carcinoma cell line, MDA MB 231 was obtained from National Centre For Cell Science (NCCS), Pune, India. The medium used for culturing the cell is MEM (Invitrogen; MEM with NEAA (non essential amino acids) and L-Glutamine) with 10% FBS

(Fetal bovine serum from HIMEDIA) and 1% antibiotic solution (Penstrep solution from HIMEDIA). The culture flask containing the cell line is kept in the CO_2 incubator with the level of CO_2 maintained at 5%. With the utilization of medium the color of the medium changes from red to orange and then pale yellow because of change in pH of the medium.
The steps for cell culture was as followed:

1. The cells were harvested first.

- Cells were grown in suspension i.e. 1 x 107 cells. The number of cells was determined. The appropriate number of cells was pelleted by centrifuging for 5 min at 300 x g in a centrifuge tube. Carefully removed all supernatant by aspiration completely from the cell culture medium.
- To trypsinize and collect cells: The number of cells was determined. The medium was aspirated, and the cells were washed with PBS. Then the PBS was aspirated, and 0.1–0.25% trypsin in PBS was added. After the cells detach from the flask, medium (containing serum to inactivate the trypsin)was added , the cells were transferred to an RNase-free glass or polypropylene centrifuge tube and centrifuged at 300 x g for 5 min. The supernatant was aspirated completely, and proceeded to step 2.

2. The cells was disrupted by adding Buffer RLT:

- For pelleted cells, loosen the cell pellet thoroughly by flicking the tube. 350 µl Buffer RLT was added. Vortexed or pipetted to mix, and ensured that no cell clumps were visible and proceeded to step 3.

3. The lysate was homogenize for 30 s using a rotor–stator homogenizer and proceeded to step 4.
4. 1 volume of 70% ethanol was added to the homogenized lysate, and mixed well by pipetting. Did not centrifuge.
5. 700 µl of each sample was transferred from step 4, including any precipitate to each RNeasy spin column on the vacuum manifold.
6. The vacuum was switched on and was applied until transfer was complete. Then switched off the vacuum and ventilated the vacuum manifold.
7. 700 µl Buffer RW1was added to each RNeasy spin column.
8. The vacuum was switched on and was applied until transfer was complete. Then switched off the vacuum and ventilated the vacuum manifold.

9. 500 µl Buffer RPE was added to each RNeasy spin column.
10. The vacuum was switched on and was applied until transfer was complete. Then switched off the vacuum and ventilated the vacuum manifold.
11. 500 µl Buffer RPE was added to each RNeasy spin column.
12. The vacuum was switched on and was applied until transfer was complete. Then switched off the vacuum and ventilated the vacuum manifold.
13. The RNeasy spin columns was removed from the vacuum manifold, and was placed each in a 2 ml collection tube. The lids were closed gently, and centrifuged at full speed for 1 min.
14. Each RNeasy spin column was placed in a new 1.5 ml collection tube. 30–50 µl RNase free water was added directly to each spin column membrane. The lids were closed gently, and centrifuged for 1 min at 8000 x g (10,000 rpm) to elute the RNA.
15. If the expected RNA yield is >30 µg, then step 15 was repeated using another 30–50 µl RNase free water or using the eluate from step 14 (if high RNA concentration is required). The collection tubes were reused from step 14.

Note: If using the eluate from step 14, the RNA yield will be 15–30% less than that obtained using a second volume of RNase-free water, but the final RNA concentration will be higher.

3.5.2. RNA ISOLATION

The kit used for RNA isolation was from QIAGEN.

1. A maximum of 1×10^7 cells was harvested, as a cell pellet or by direct lysis of/in vessel. The appropriate volume of Buffer RLT was added.
2. 1 volume of 70% ethanol was added to the lysates and mixed well by pipetting. Did not centrifuge. Proceeded immediately to step 3.
3. Up to 700 µl of the sample was transferred, including any precipitation, to an RNeasy Mini spin column placed in a 2ml collection tube (supplied). The lid was closed and centrifuged for 15s at ≥8000×g. The flow –through was discarded.
4. 700 µl Buffer RW1 was added to the RNeasy spin column. The lid was closed and centrifuged for 15s at 8000×g. The flow –through was discarded.
5. 500 µl Buffer RPE was added to the RNeasy spin column. The lid was closed and centrifuged for 15s at ≥8000×g. The flow –through was discarded.
6. 500 µl Buffer RPE was added to the RNeasy spin column. The lid was closed and centrifuge for 2 min at ≥8000×g.

7. The RNeasy spin column was placed in the new 1.5 ml collection tube. 30-50 µl RNase- free water was added directly to the spin column membrane. The lid was closed and centrifuged for 1min at ≥8000×g to elute the RNA.
8. If the expected RNA yield is >30 µg, then step 7 was repeated using another 30-50µl of RNase- free water, or using the eluate from step-7. The collection tubes were reused from step-7.
9. The purity and yield of RNA yield was measured by **Eppendorf NanoDrop**. It is a cuvette free spectrophotometer which eliminates the need for other sample containment devices and allows for clean up in seconds. It measures 1 µl samples with high accuracy and reproducibility. The full spectrum (220nm-750nm) spectrophotometer utilizes a patented sample retention technology that employs surface tension alone to hold the sample in place. A 1 µl sample is pipetted onto the end of a fiber optic cable (the receiving fiber). A second fiber optic cable (the source fiber) is then brought into contact with the liquid sample causing the liquid to bridge the gap between the fiber optic ends. The gap is controlled to both 1mm and 0.2 mm paths. A pulsed xenon flash lamp provides the light source and a spectrometer utilizing a linear CCD array is used to analyze the light after passing through the sample. The instrument is controlled by PC based software, and the data is logged in an archive file on the PC.

3.5.3. cDNA Synthesis

cDNA synthesis was carried out using SuperScript First-Strand Synthesis System for RT-PCR by Invitrogen using oligo dT primers.

The steps in cDNA synthesis:

1. Each of the components were mixed and briefly centrifuged before use.
2. For each reaction, the following in a sterile 0.2 or 0.5ml tube was combined.

Components	Amount
RNA	4 µl
10 mM dNTP mix	1 µl
Primer (0.5µg/µl oligo $(dT)_{12-18}$ or 2µM gene specific primer)	1µl
DEPC treated water	4µl

1. The RNA/primer mixture at 65·c for 5 minutes was incubated, and then placed on ice for at least 1 minute.
2. In a separation tube, the following 2X reaction was prepared by adding each component in the indicated order.

Components	1RXn	10 RXns
10X RT buffer	2 µl	20 µl
25mM $Mgcl_2$	4 µl	40 µl
0.1M DTT	2 µl	20 µl
RNase out ^{TM} (400/ µl)	1 µl	10 µl

1. 9µl of the 2X reaction mixture was added to each RNA/primer mixture from step3, mixed gently and collected by briefly centrifuge.
2. It was incubate at 42°c for 2 minutes.
3. 1µl of super script TM II RT was added to each tube.
4. It was incubate at 42°c for 50 minutes.
5. The reaction was terminated at 70·c for 15 minutes. Chilled on ice.
6. The reaction was collected by brief centrifugation. 1µl of RNase H was added to each tube and incubated for 20minutes at 37·c. The reaction was used for PCR immediately.

3.5.4. Real time PCR

Real time PCR is a method that allows exponential amplification of DNA sequences and simultaneously quantifies it. This system is based on the detection and quantitation of a

fluorescent probes. Probes which are used in qRT-PCR are taqman probes, molecular beacon , SYBR® Green, displacing probes, light up probes etc. For present study we used SYBR® Green probe which is a frequently used fluorescent DNA binding probe and relies on the sequence specific detection dye. The genes which are taken for the experiment, are of two types test genes and reference genes. Reference genes are used to check for the expression of test genes i.e. how much fold they have increased or decreased with respect to normal expression. Reference genes should have following feature: the standard gene should have the same copy number in all cells, it should be expressed in all cells, a medium copy number is advantageous since the correction should be more accurate. For this experiment the test genes were ABCA8 and SMC4 and reference gene used is beta-actin. Beta actin is a house keeping gene also called as constitutive genes which are required for the maintenance of fundamental cellular function, and are expressed in all cells of an organism under normal and patho-physiological conditions. The total reaction volume could either be 5µl or 10µl. We prepared a total reaction volume of 10µl.

Calculation:

We took 3 genes, therefore,

3 genes x 3 replicates = 9 x 10 = 90µl

=100µl (in case of pipetting error do addional 10 µl is taken)

i) **SYBR ® Green master mix:-**

The stock solution is 2X concentration and working solution of 1 X concentration is prepared.

2X x (? µl) = 1X x 100µl

⇨ (? µl) = 1X x 100µl / 2X

= 50µl

SYBR ® Green master mix= 50µl

ii) **cDNA:**

cDNA = cDNA stock : distilled water

1 : 20

cDNA=3µl

Distilled water = 57µl

Total= 3+57 µl=60µl

For each reaction we require 4µl of cDNA

Therefore, for 10 reaction= 4x10= 40µl

From the above made 60 µl we took 40µl

iii) **Primer:-** The stock solution contains 10µM, we require 500nM for each reaction

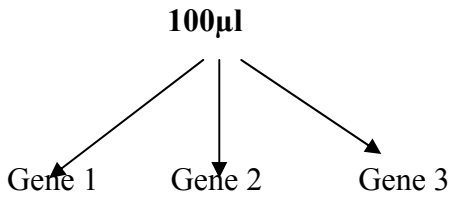
10 µM x (? µl) = 500x1000µMx100

⇨ (? µl) = 500x1000µMx100 / 10 µM

⇨ = 5µl

For forward and reverse primer 5x2=10µl

SYBR ® Green master mix + cDNA + Primer= (50 +40+10) µl= 100 µl



For each gene 3 replicates is taken

- For Real time PCR analysis mainly 2 types of primers are used.
- There are 3 genes taken. One is EGR1 gene, another is LAMB1 gene and another is normal gene for checking the expression.
- The primer sequence are ordered from Sigma and sequence for each gene are**:**

Table 2: Table showing the sequence of the forward and backward primers

PRIMER	**TYPE**	**SEQUENCE**
EGR1	***Forward***	GGTCAGTGGCCTAGTGAGC
	Reverse	GTGCCGCTGAGTAAATGGGA
LAMB1	***Forward***	AGGAACCCGAGTTCAGCTAC
	Reverse	CACGTCGAGGTCACCGAAAG
β actin	***Forward***	CATGTACGTTGCTATCCAGGC
	Reverse	CTCCTTAATGTCACGCACGAT

- Real-time PCR was carried out in Eppendorf Masterplex Real Time PCR.

1. The primer concentrations were normalized and gene-specific forward and reverse primer

pair was mixed. Each primer (forward or reverse) concentration in the mixture was 3.5 µl.

2. The experiment was set up and the following PCR program was made on. A copy of the setup file was saved and all PCR cycles were deleted. The threshold frequency taken was 33%. The cycle temperatures taken were as follows:

Table 3: Cycle temperature and time for qRT-PCR

STAGE	TEMPERATURE (°C)	TIME	CYCLE
Stage 1	95	20 sec	1
Stage 2	95 55 68	15 sec 15 sec 20 sec	40
Stage 3	95 60 95	15 sec 15 sec 15 sec	1

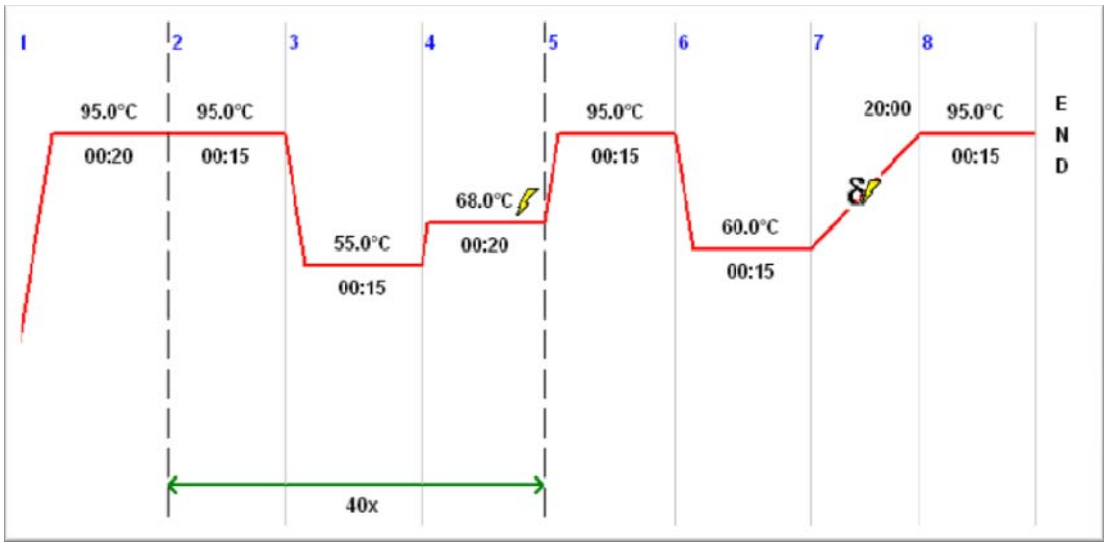


Figure 06. Cycle temperature and time for qRT-PCR

Results and Discussion

4. RESULTS & DISCUSSIONS

In study of breast cancer cells, EGR1 and LAMB1 gene expression was up regulated compared with the normal cells. The observation was confirmed by doing experiment Insilco as well as in lab. Experiment was performed by RNA isolation from MDA Mb 231 breast cancer cells and its expression was checked by running real time PCR and agarose gel.β actin a house keeping gene that shows same expression in all kind of cells was visualized for its expression. When normal cell was compared with breast cancer cells EGR1 expression was down in normal cells while well bright band was observed in breast cancer cells suggesting it is up regulated in breast cancer and LAMB1 expression is up regulated.

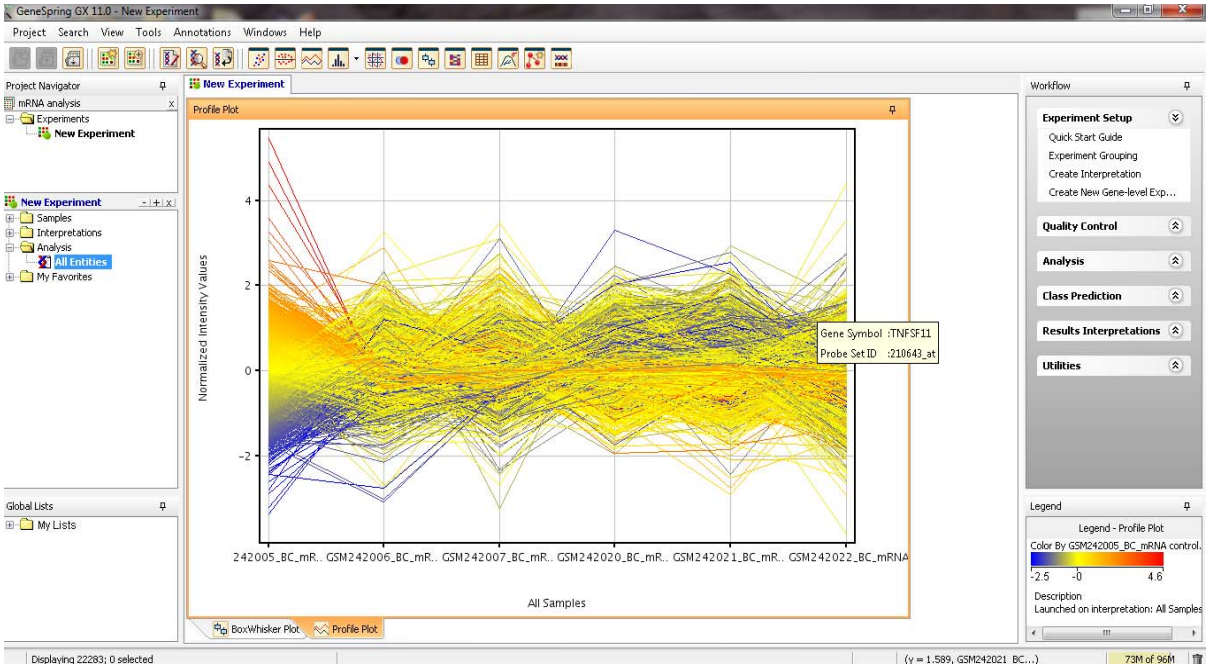


Figure 07: Intensity map of the control and test samples

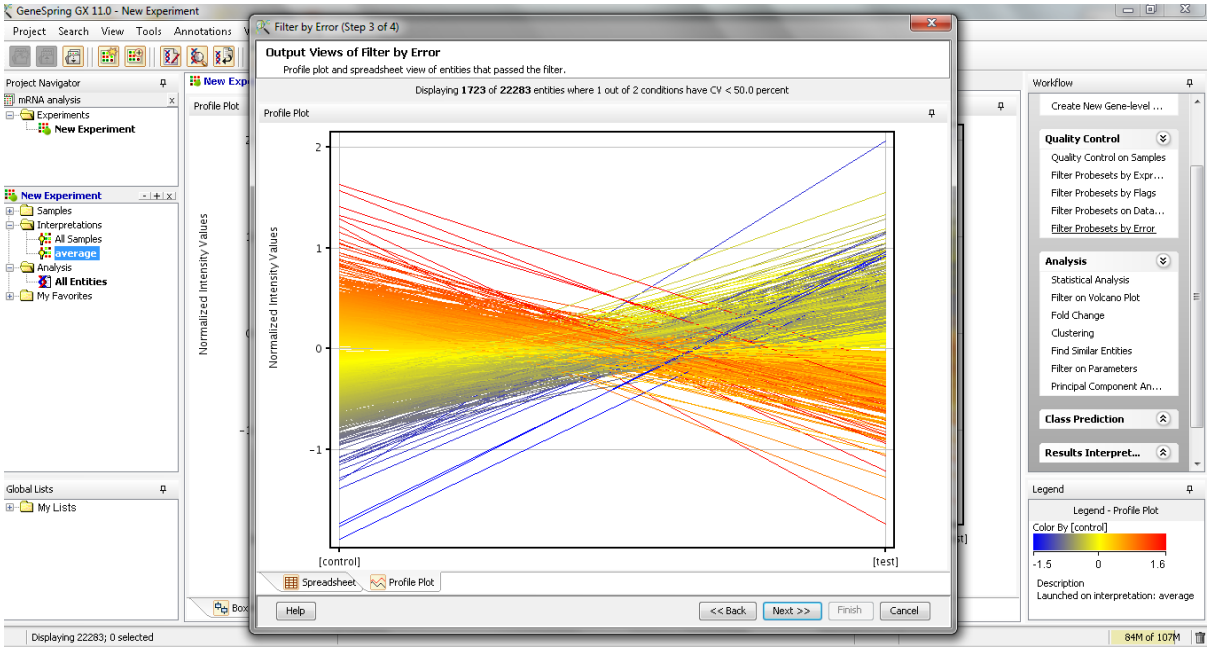


Figure 08: Experimental grouping is done by add parameters to Average

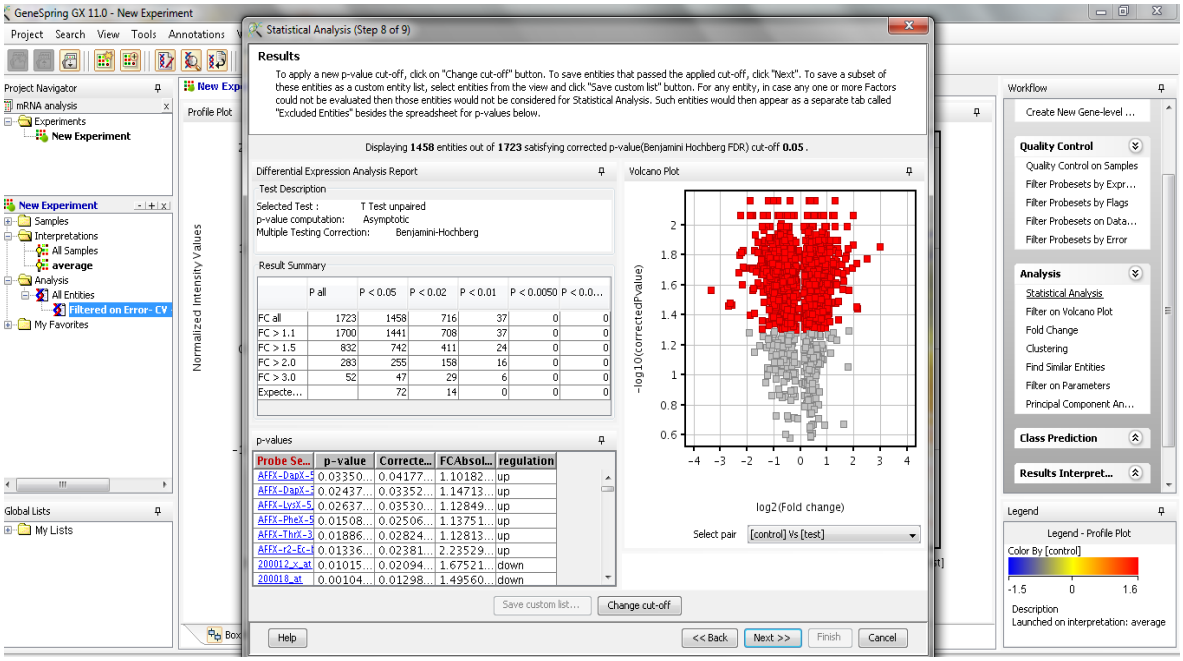


Figure 09: Filtering Probe sets by errors

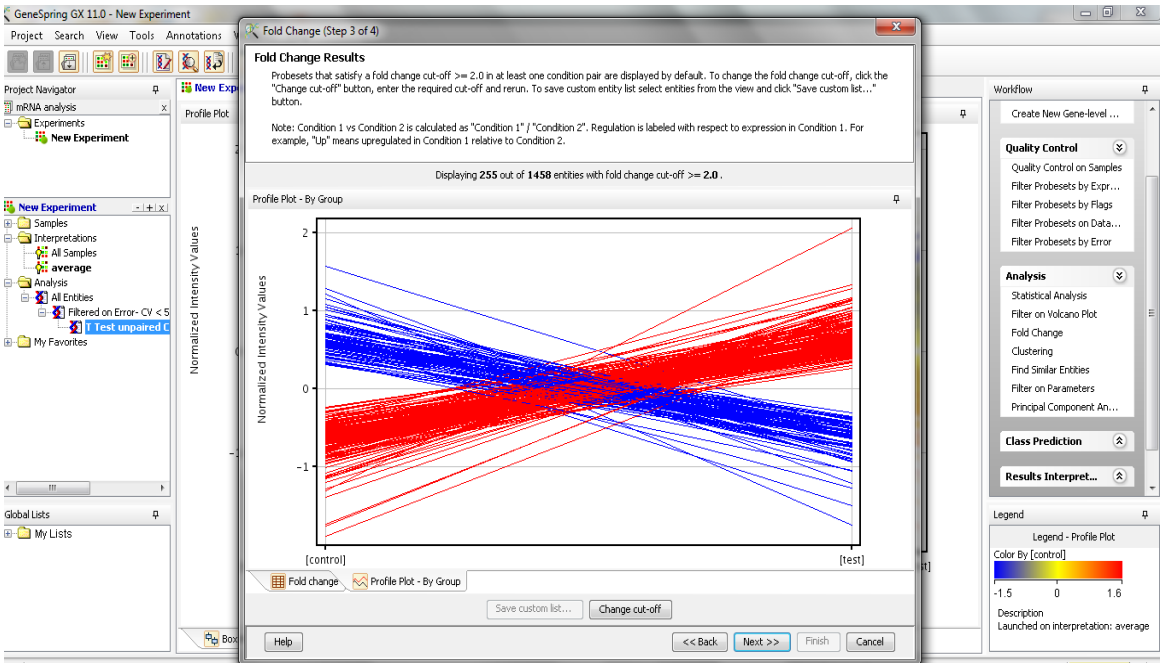


Figure 10: T-test is unpaired chosen for 2 sets of data

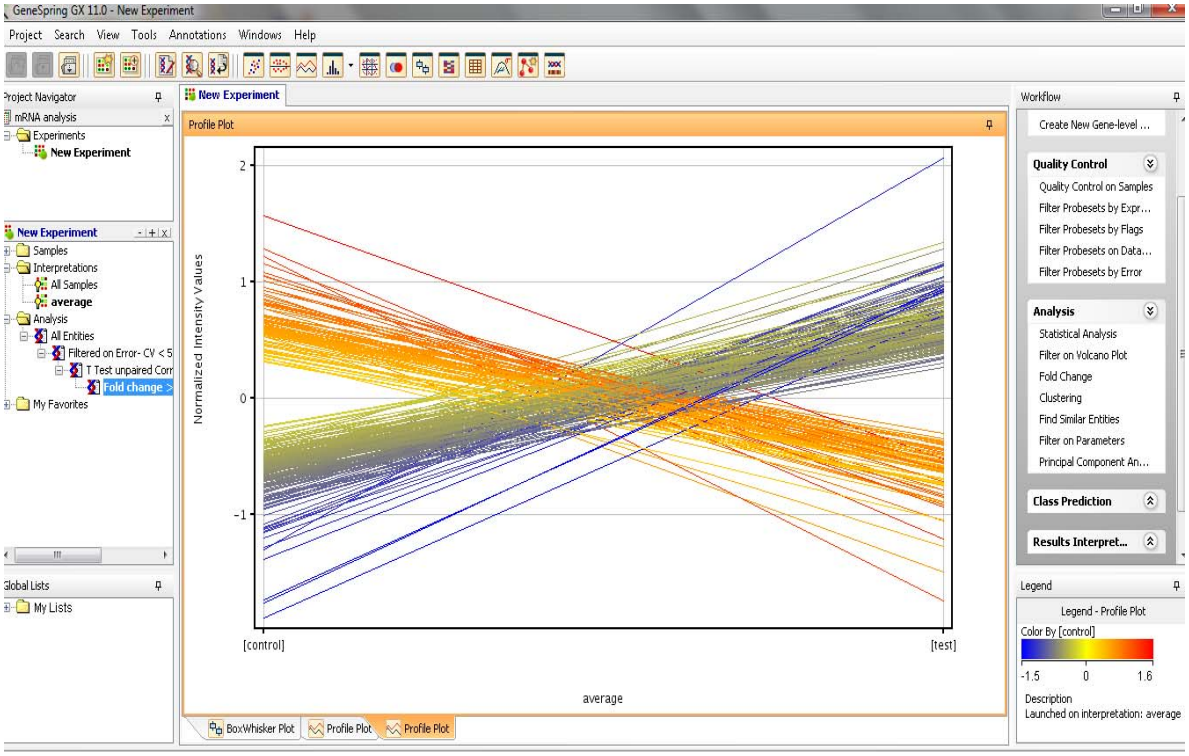


Figure 11: Fold change results by taking cut-off >=2.0.

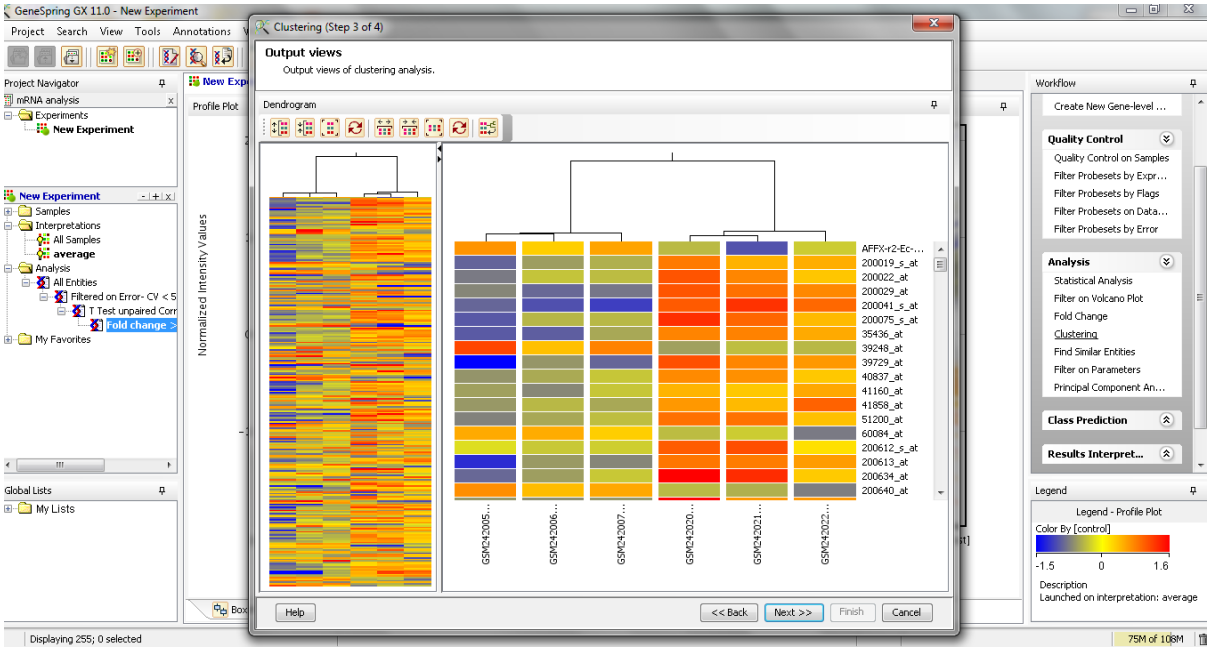


Figure 12: Hierarchical clustering of output views.

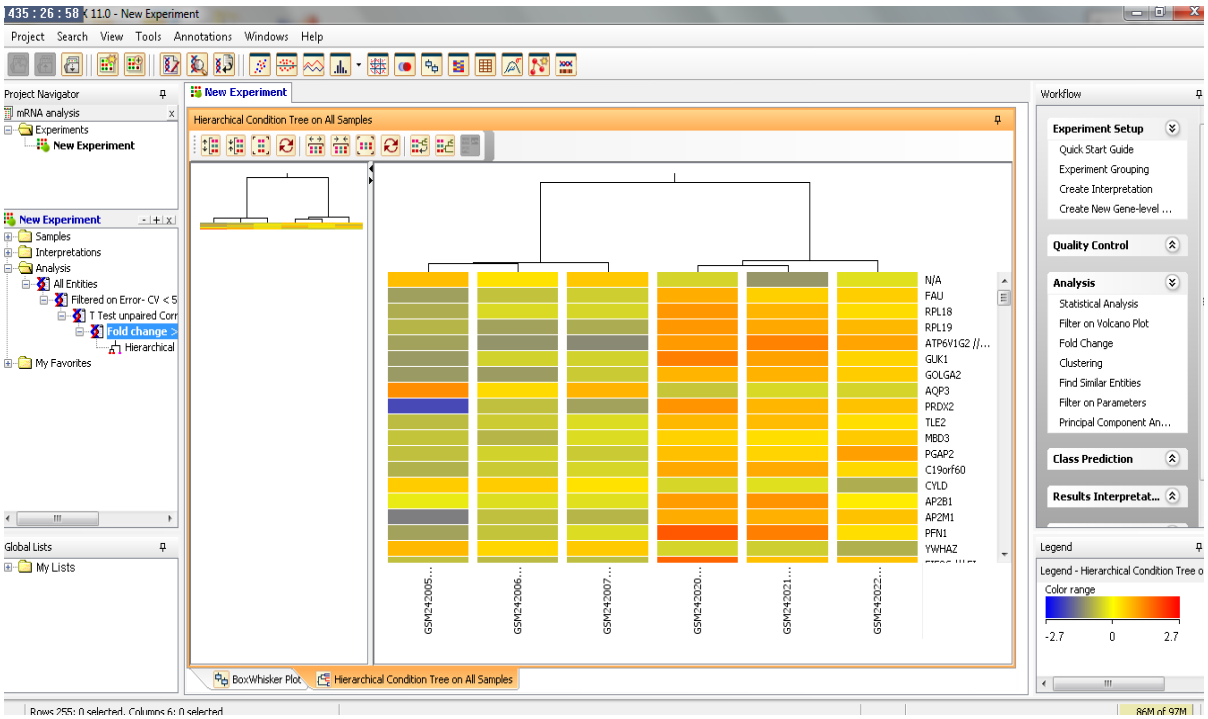


Figure 13: Hierarchical clustering by taking Interpretation all samples

- A list of 255 genes is generated by gene spring analysis that is differentially expressed in breast cancer patients. Out of 255, 153 genes were reported to be down-regulated and 102 genes were up-regulated in breast cancer.
- These genes are further analyzed in **Cluster 3** and **java treeview** software which generated a hierarchical clustering dendrogram as follows.

Cluter3 and java treeview result:-

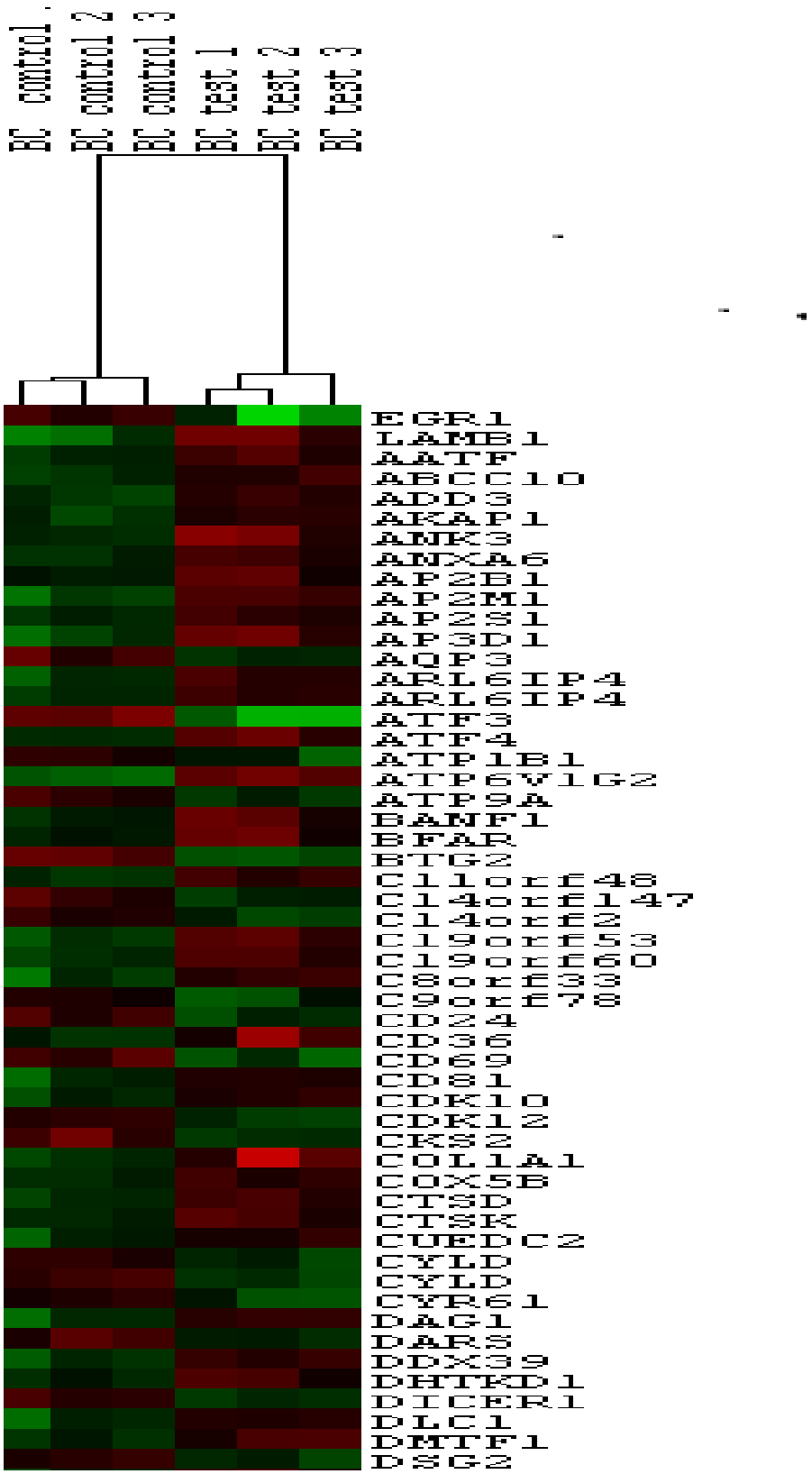


Figure 14: Clustering results of genes using java treeview software

- ❖ From the list of 255 genes which are obtained from using GeneSpring software were further analyzed by using different software and databases.
- ❖ Genes which have fold change value >3 are taken for analysis from the list of 255 genes.
- ❖ It is done by searching in different data base and literature search.
- ❖ From that 2 genes are selected for further analysis by seeing its regulation in different pathways.
- ❖ These 2 genes are EGR1 (Early growth response 1) and LAMB1 (Laminin beta 1).

- These 2 genes are analyzed in String database and Genomatrix software for to the interaction with the genes which play an important role in cancer invasiveness.

String database result

The interaction of EGR1 and LAMB1 gene with other genes are documented from string database.

Association of EGR1 gene with other gene

- TAKING EXPERIMENT-DATABASE-TEXTMINING PARAMETER ANALYSIS

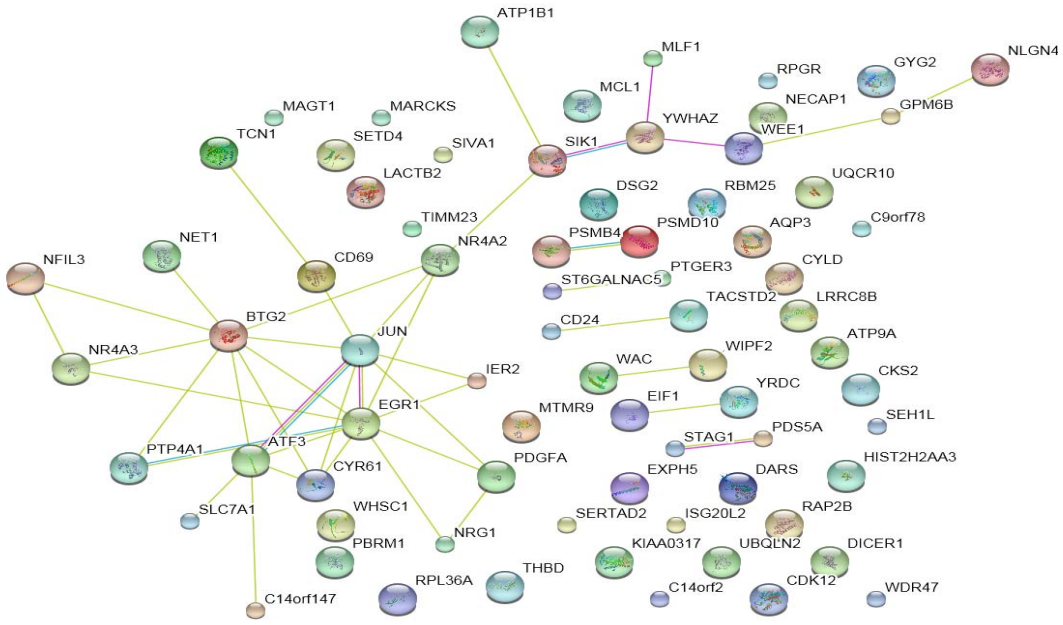


Figure 15: Interaction of EGR1 gene with other gene by taking Experiment-Database-Text mining parameter analysis

- **TAKING TEXTMINING PARAMETER ONLY**

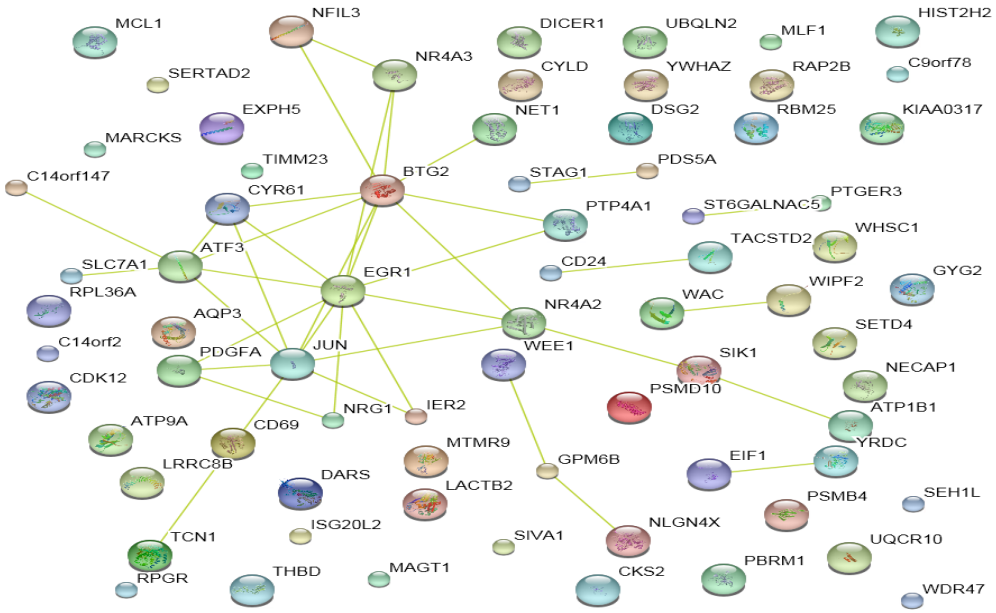


Figure 16: Interaction of EGR1 gene with other gene by taking Text mining parameter analysis

- **TAKING EXPERIMENT-TEXTMINING PARAMETER ONLY**

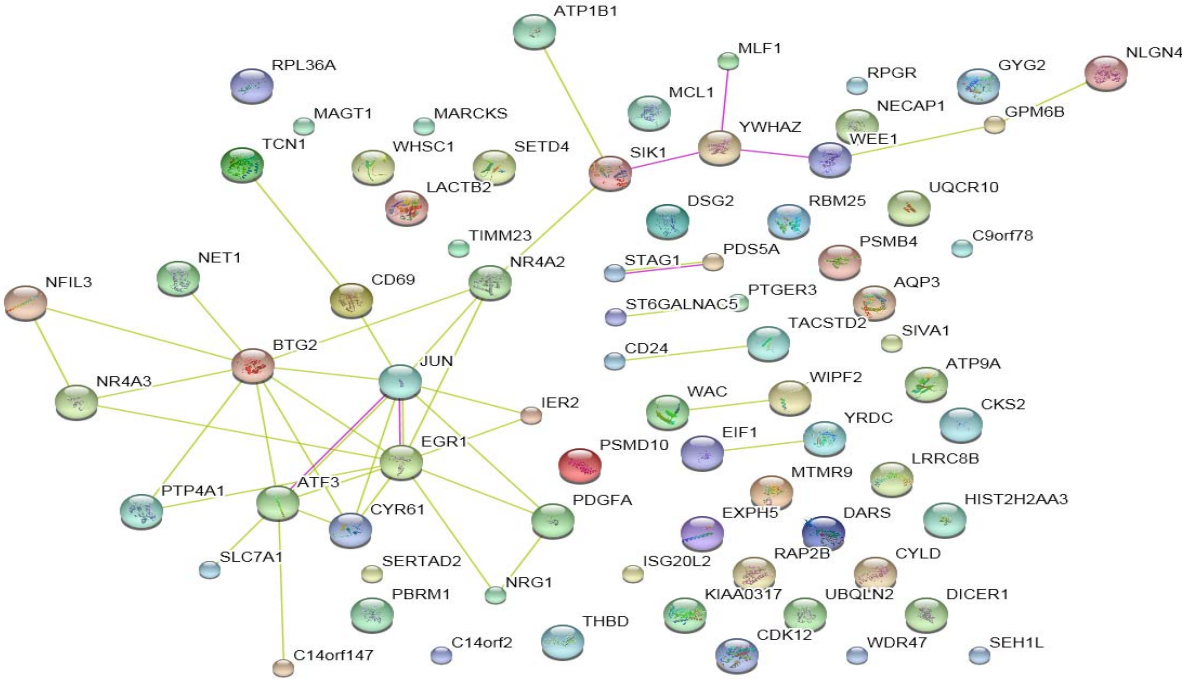


Figure 17: Interaction of EGR1 gene with other gene by taking Experiment-Text mining parameter analysis

- **TAKING DATABASE-TEXTMINING PARAMETER ONLY**

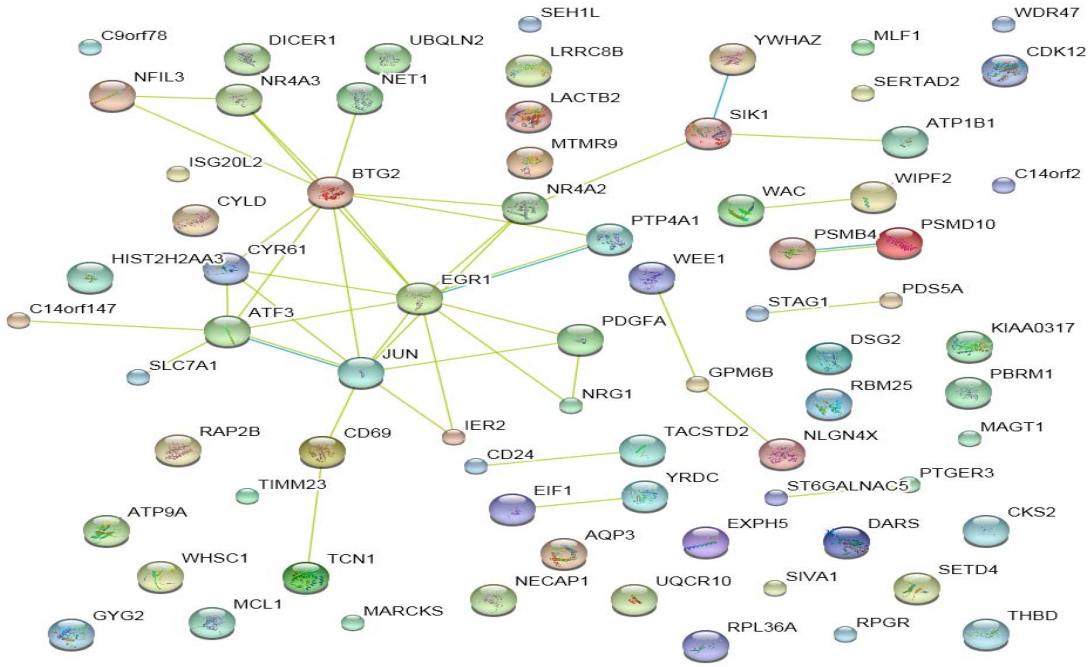


Figure 18: Interaction of EGR1 gene with other gene by taking Database-Text mining parameter analysis

Association of LAMB1 gene with other gene

- TAKING EXPERIMENT-DATABASE-TEXTMINING PARAMETER ANALYSIS

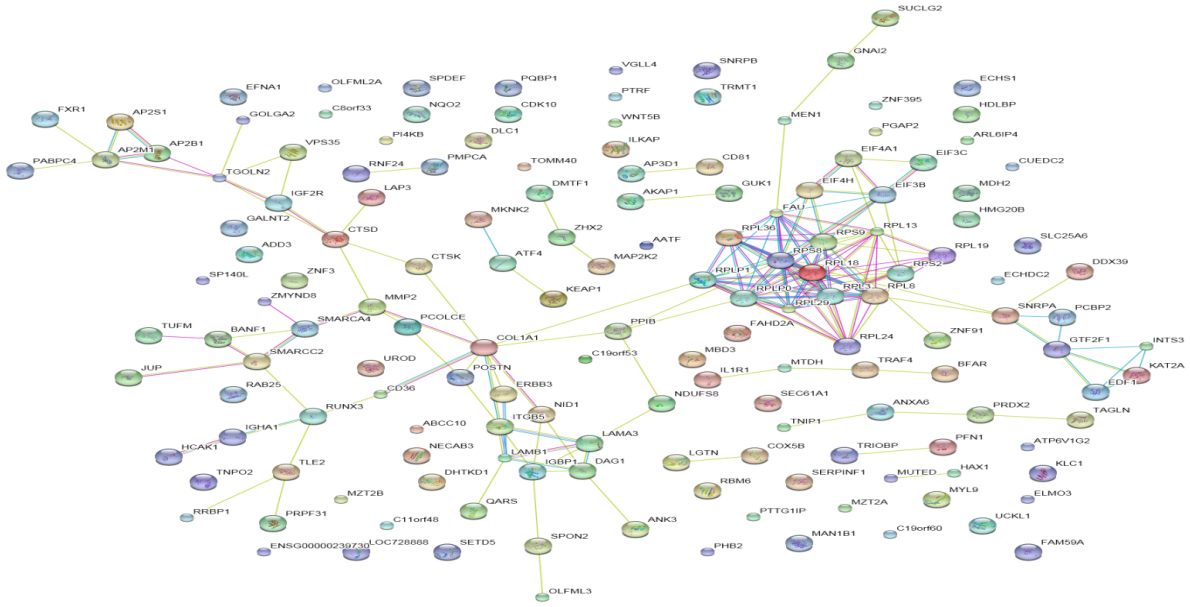


Figure 19: Interaction of LAMB1 gene with other gene by taking Experiment-Database-Text mining parameter analysis

- **TAKING EXPERIMENT-DATABASE PARAMETER ONLY**

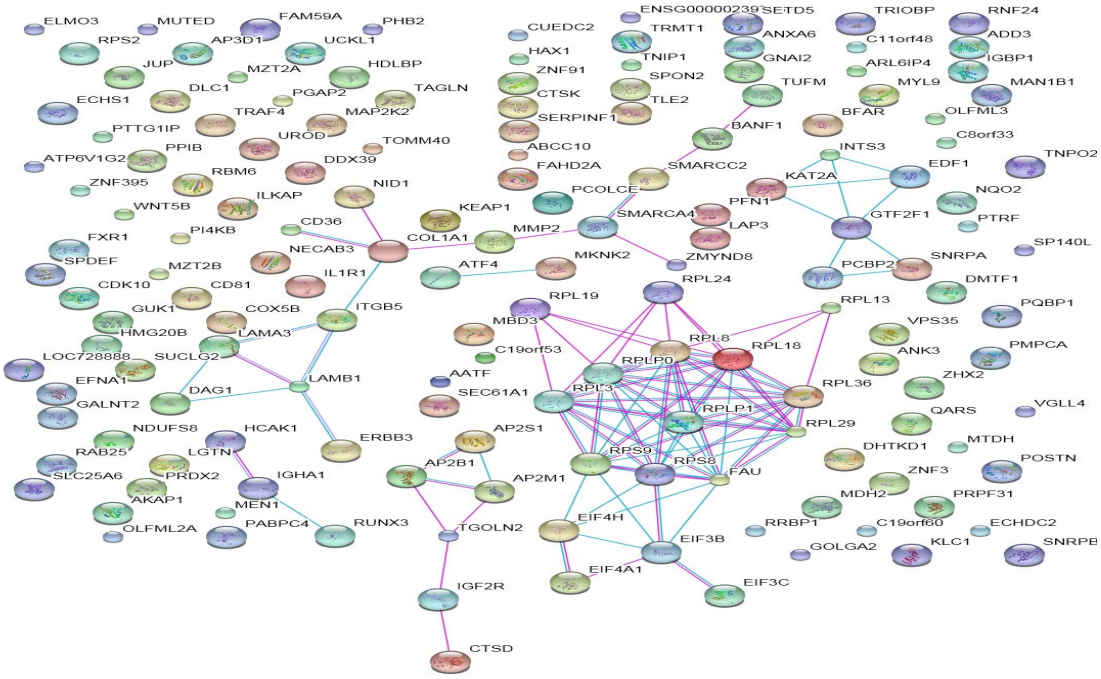


Figure 20: Interaction of LAMB1 gene with other gene by taking Experiment-Database parameter analysis

- **TAKING EXPERIMENT-TEXTMINING PARAMETER ONLY**

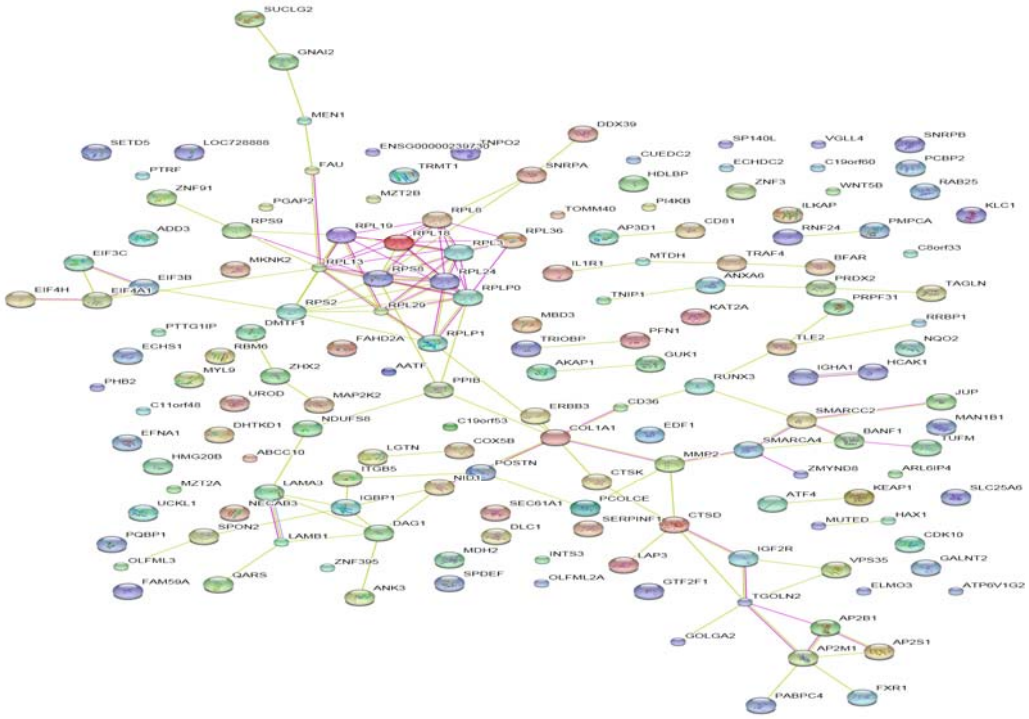


Figure 21: Interaction of LAMB1 gene with other gene by taking Experiment-Text mining parameter analysis

- **TAKING DATABASE-TEXTMINING PARAMETER ONLY**

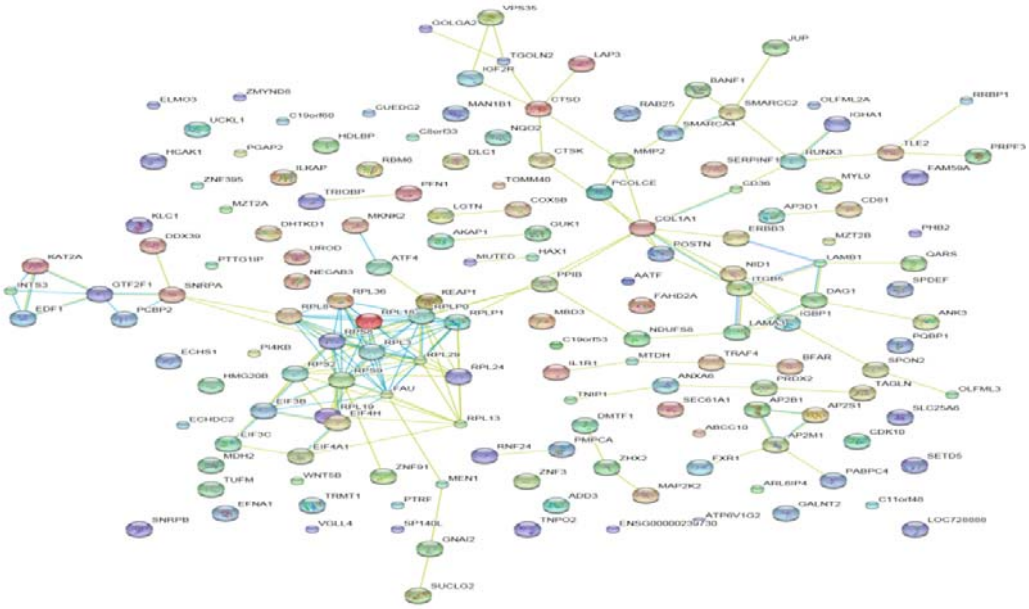


Figure 22: Interaction of LAMB1 gene with other gene by taking Database-Text mining parameter analysis

- Then these 2 genes are also analyzed in genomatrix software to find association with its neighboring gene.

Genomatrix results

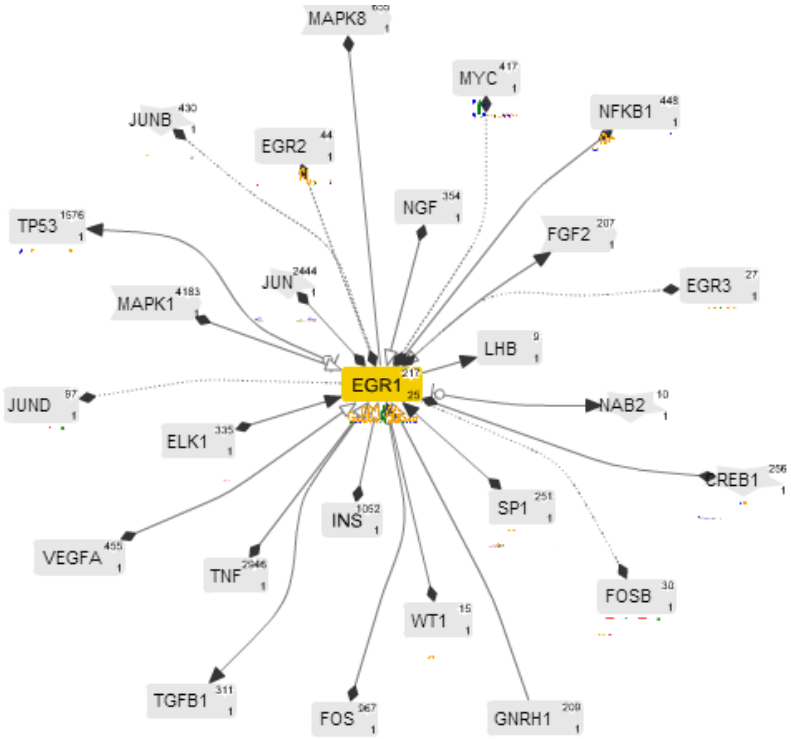


Figure 23: Egr1showing Association with its neighboring gene

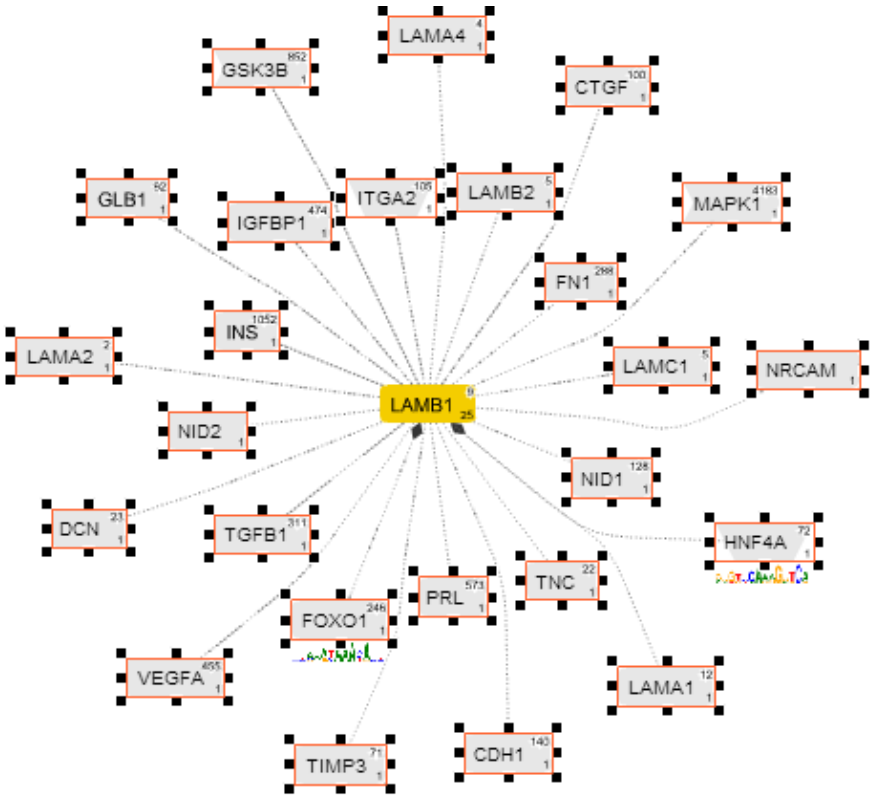


Figure 24: Lamb1 showing association with its neighboring gene

- These are the symbols by which we can know which type of association occurring in between the genes.

2 genes are associated by co-citation.

2 genes are associated by expert-curation.

Gene A activates Gene B.

Gene A inhibits Gene B.

Gene A modulates Gene B.

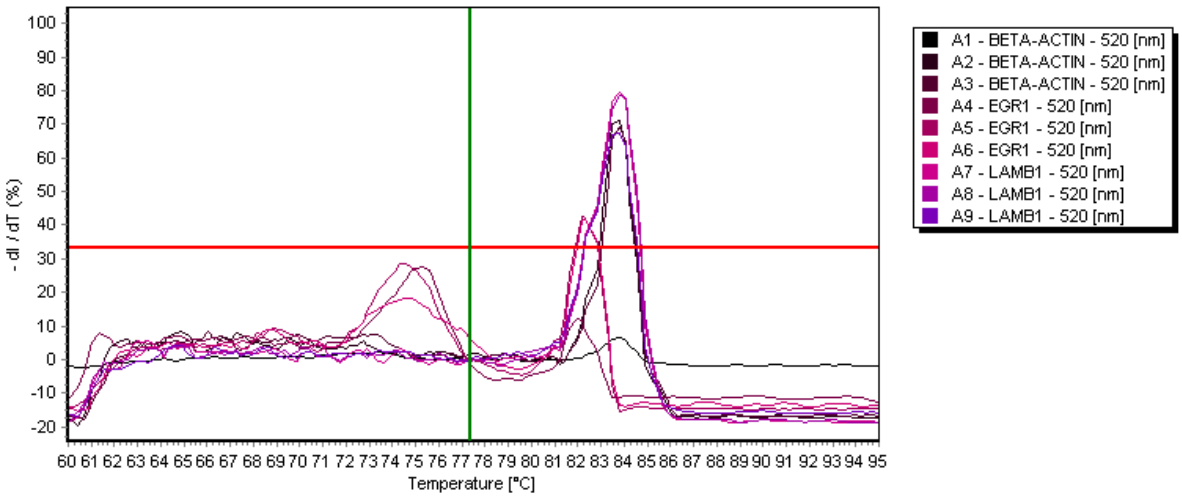
Gene A alters the state of Gene B.

●/◆/▶ If gene A has a known TF binding site matrix and gene B has a corresponding **binding site** in one of its promoters the arrow is filled black. For interactions that involve a complex, this arrow type is never used. To look for promoter bindings in this case, double-click on the edge and select the interaction of interest.

o/◇/▷ There is no promoter binding noted

Real time PCR results

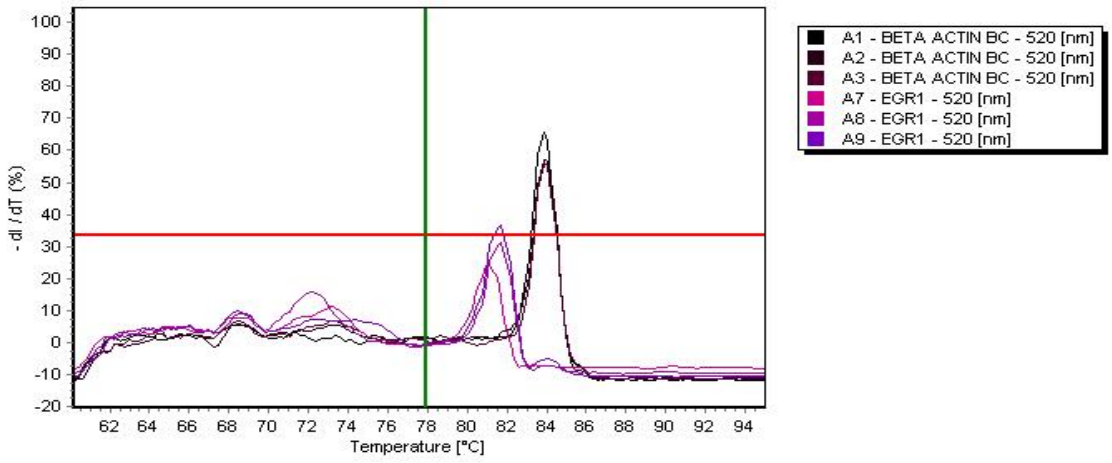
Melting curve analysis of EGR1 and LAMB1 with β actin



Threshold: 33%

Figure 25:Showing the results of melting curve analysis of both EGR1 and LAMB1 using β actin as standard

Melting curve analysis of only EGR1 gene with β actin



Threshold: 33%

Figure 26:Showing the results of melting curve analysis of EGR1 gene using β actin as standard

In many tumor types, such as glioblastoma, lymphoma, and carcinomas of the lung and breast, EGR1 is low or absent, (Levin et.al, 1995; Huang et.al.1997; Calogero et.al, 2001 and Joslin et.al, 2007) indicating a tumor suppressive role. From the in silico data it was stated that Egr1 gene expression is down regulated in breast cancer. But from the melting curve analysis through RTPCR in breast cancer it was seen that it is up regulated. In other cancers like prostate cancer EGR1 has also been reported up regulated (Gregg et.al, 2011) .

Melting curve analysis of only LAMB1 gene with β actin

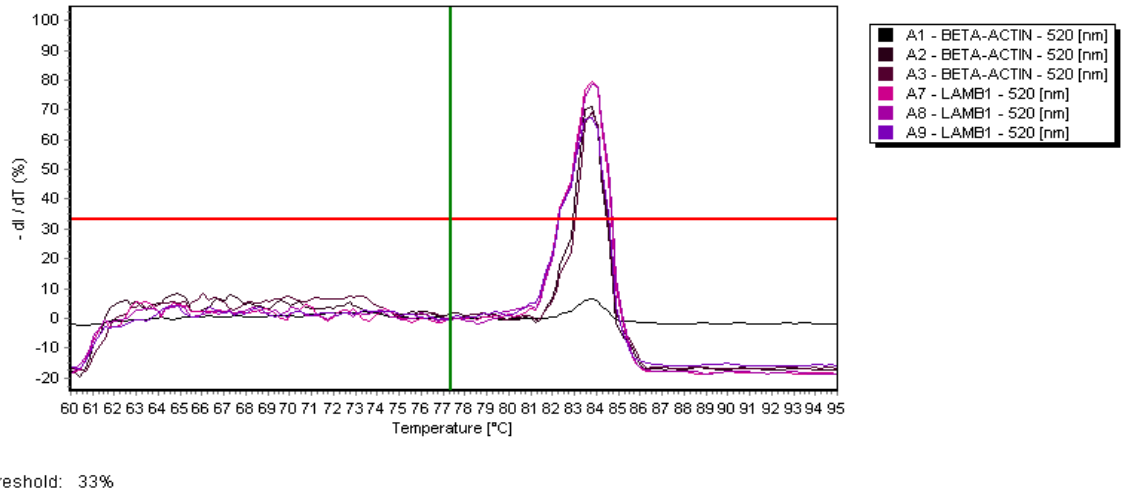


Figure 27:Showing the results of melting curve analysis of LAMB1 gene using β actin as standard

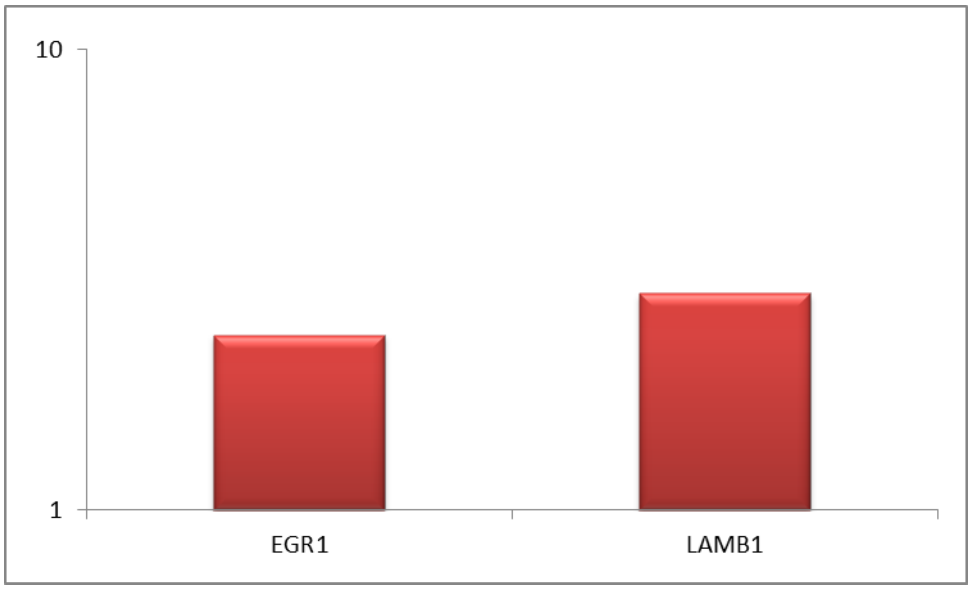


Figure 28: Relative expression of EGR1 AND LAMB1 with respect to control

It was observed that EGR1 down regulated and LAMB1 was up expressed in breast cancer. So it may concurred that a coordinated endeavor between basic molecular biology functional genomics and proteomics and cell signaling may provide information on the relationship of this down and up regulation of genes .

Conclusion

5. CONCLUSION

We analyzed microarray data of breast cancer tissues samples and compared with array of normal breast epithelium to obtain a list of 255 genes that are differentially expressed in breast cancer patients. Out of 255, 153 genes were reported to be down-regulated and 102 genes were up-regulated in breast cancer. From this differentially expressed gene list, we choose two genes, EGR1 & LAMB1 for experimental validation, because EGR1 is a transcription factor whereas LAMB1 is an extracellular matrix glycol protein. And they have significant role in tumor suppression and metastasis. Their de-regulation might be contributing factor of breast cancer by affecting various pathways/processes in the body. The expressions of these two genes were validated by qRT-PCR in MDA-MB-231 breast cancer cell lines, followed by elucidating their association in various biological networks. We found EGR1 to be down-regulated in breast cancer and this de-regulation possibly affect other downstream genes in pathways and networks related to breast cancer. Further analysis encompassing this gene might provide deeper insight towards understanding breast cancer aetiology and give clue towards manipulating the associated pathways for cure of this disease. LAMB1 expression was up-regulated in case of breast cancer as compared to normal breast sample and further study may hint towards understanding mechanism metastasis in breast cancer.

Future scope

Future studies on genes out of 255 differentially expressed genes which are yet not been validated in breast cancer, but are expressed in breast cancer will provide deeper insight into disease etiology of breast cancer.

References

6. REFERENCES

1. Fey MF. Impact of the Human Genome Project on the clinical management of sporadic cancers. Lancet Oncol. 2002 Jun;3(6):349-56.

2. Kwiatkowski P, Wierzbicki P, Kmieć A, Godlewski J. DNA microarray-based gene expression profiling in diagnosis, assessing prognosis and predicting response to therapy in colorectal cancer. Postepy Hig Med Dosw (Online). 2012 Jun 11;66:330-8.

3. Bertucci F, Houlgatte R, Nguyen C, Viens P, Jordan BR, Birnbaum D. Gene expression profiling of cancer by use of DNA arrays: how far from the clinic? Lancet Oncol. 2001 Nov;2(11):674-82.

4. Zajchowski DA, Bartholdi MF, Gong Y, Webster L, Liu HL, Munishkin A, Beauheim C, Harvey S, Ethier SP, Johnson PH. Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells. Cancer Res. 2001 Jul 1;61(13):5168-78.

5.Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52.

6.Brennan DJ, O'Brien SL, Fagan A, Culhane AC, Higgins DG, Duffy MJ, Gallagher WM. Application of DNA microarray technology in determining breast cancer prognosis and therapeutic response. Expert Opin Biol Ther. 2005 Aug;5(8):1069-83.

7. Kleinsmith L.J, Principles of cancer biology,1-9,45,2009

8.Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol. 1998 Jul;16(7):301-6.

9. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31;415(6871):530-6.

10. Mallick B.Ghosh J; Bioinformatics: principles and application,77-79,2012 11.

11. Pattin KA, Moore JH. Role for protein-protein interaction databases in human genetics. Expert Rev Proteomics. 2009 Dec;6(6):647-59.

12. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010 Mar;31(3):115-23.

13. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM,Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007 Mar;25(3):309-16.

14. Wang PI, Marcotte EM. It's the machine that matters: Predicting gene function and phenotype from protein networks. J Proteomics. 2010 Oct 10;73(11):2277-89.doi: 10.1016/j.jprot.2010.07.005. Epub 2010 Jul 15.

15. Levin WJ, Press MF, Gaynor RB, Sukhatme VP, Boone TC, Reissmann PT, Figlin RA,Holmes EC, Souza LM, Slamon DJ. Expression patterns of immediate earlytranscription factors in human non-small cell lung cancer. The Lung Cancer Study Group. Oncogene. 1995 Oct 5;11(7):1261-9.

16. Joslin JM, Fernald AA, Tennant TR, Davis EM, Kogan SC, Anastasi J, CrispinoJD, Le Beau MM. Haploinsufficiency of EGR1, a candidate gene in the del(5q), leads to the development of myeloid disorders. Blood. 2007 Jul 15;110(2):719-26.

17. Huang RP, Fan Y, de Belle I, Niemeyer C, Gottardis MM, Mercola D, Adamson ED. Decreased Egr-1 expression in human, mouse and rat mammary cells and tissues correlates with tumor formation. Int J Cancer. 1997 Jul 3;72(1):102-9.

18. Calogero A, Arcella A, De Gregorio G, Porcellini A, Mercola D, Liu C, Lombari V, Zani M, Giannini G, Gagliardi FM, Caruso R, Gulino A, Frati L, Ragona G. The early growth response gene EGR-1 behaves as a suppressor gene that is down-regulated independent of ARF/Mdm2 but not p53 alterations in fresh human gliomas. Clin Cancer Res. 2001 Sep;7(9):2788-96.

19. Petz M, Them NC, Huber H, Mikulits W. PDGF enhances IRES-mediated translation of Laminin B1 by cytoplasmic accumulation of La during epithelial to mesenchymal transition. Nucleic Acids Res. 2012 Oct;40(19):9738-49.

20. Petz M, Them N, Huber H, Beug H, Mikulits W. La enhances IRES-mediated translation of laminin B1 during malignant epithelial to mesenchymal transition. Nucleic Acids Res. 2012 Jan;40(1):290-302.